ISSN 2616-5775 Vol. 5, Issue 10: 28-34, DOI: 10.25236/AJCIS.2022.051005

Prediction Analysis of Shenzhen GDP Based on ARIMA Model and Implementation in R Language

Hongye Cai, Wenxuan Qiu

College of Mathematics and Statistic, Shenzhen University, Shenzhen, 518060, China

Abstract: *GDP can measure the development level and economic situation of a region to a certain extent. This paper selects the GDP data of Shenzhen from 1980 to 2020, and constructs the Arima(2, 2, 3) model to predict the GDP data in the next 5 years. First, after Box-Cox transformation and difference operation are performed on the selected data, the Augmented Dickey-Fuller (ADF) stationarity test is carried out. Under the condition of passing the ADF test, this paper uses the maximum likelihood method to iteratively estimate the parameters of the model based on the initial value of the least squares' estimation, and then performs a mixed test on the residual data to verify the validity of the parameters. Finally, the h steps forward method is used to predict the GDP data of Shenzhen in the next five years, and the corresponding confidence interval is obtained. The relative error between the prediction results in 2021 and the actual data is 2.9%, indicating that the prediction results of the model are good and have certain feasibility.*

Keywords: *Shenzhen GDP; ARIMA model; R Studio; Forecast results*

1. Introduction

In recent years, many scholars have discussed the economic situation of Shenzhen from various aspects such as economy, medicine and environment. Guo Zhiwu et al. (2009) [1] eliminated the influence of the Spring Festival factor on the time series through the seasonal product model; Zheng Huimin et al. (2016) [2] studied the incidence trend of infectious diseases in Shenzhen, which is of great practical significance; Yan Zhouning et al. (2018)[3] analyzed the time series of atmospheric PM2.5 concentration in Shenzhen, which plays an important role in seasonal environmental management; Yi Zhiguo et al. (2021)[4] predicted cigarette sales in the next year , which provides a certain reference for the reasonable launch of the market.

Other scholars are also working on future forecasts of GDP. Chen Congcong selected the GDP data of Shandong Province from 1975 to 2013, established the ARIMAX model, and obtained a short-term macro forecast [5]; Yan Yanwen analyzed the GDP data of Shandong Province from 1975 to 2015, and established ARIMA (1,1,1) model, the results show that the prediction effect is good [6] ;Zhao Zimeng selected 1991-2016 GDP data for analysis, and concluded that Chengdu's economy will be in a high growth stage in the next few years [7]; Xu Mingyan used the ARIMA model, the BP neural network model, the combined model and the improved combined model to predict the GDP data of Jiangsu Province from 1970 to 2018. By comparing the relative errors of different models, she drew the conclusion that the improved combination model has better short-term prediction, and the ARIMA model is more suitable for long-term prediction [8]; Qu Haiqing et al. established the ARIMA (0,2,3) model, which reflected the trend of GDP development in Hubei Province and made short-term forecasts [9].

A large number of existing studies have proved the practicability of the ARIMA model. At the same time, the ARIMA model only needs endogenous variables without other exogenous variables, and the estimation is relatively simple. The prediction of variable data has high accuracy and reliable prediction value. Therefore, the ARIMA model will be used in this paper to predict the time series data of Shenzhen's GDP, and to study the development prospects and laws of Shenzhen through stationarity test, model order determination, residual error test, and data prediction. The final forecast of GDP data in 2021 has an error of only 2.9% from the actual value, indicating that the model has a certain reference value.

ISSN 2616-5775 Vol. 5, Issue 10: 28-34, DOI: 10.25236/AJCIS.2022.051005

2. Introducing of ARIMA

2.1 Overview of Model

The following models have been centralized, if not, replace x_t with x_t-μ. (μ is the mean)

Use $x_1, x_2, \cdots, x_n$ or $\{x_t, t=1, 2, \cdots, n\}$ represents the n observations of the time series, and denote$\{\varepsilon_t\}$ as the white noise sequence. That is,$\{\varepsilon_t\}$satisfies $E(\varepsilon_t)=0$, $Var(\varepsilon_t)=\sigma_\varepsilon^2$. The delay operator ∇^d and the difference operatorBare introduced to simplify the expression. Their definitions are:

$$Bx_t=x_{t-1} \tag{1}$$

$$\nabla^d x_t=(1-B)^d x_t \tag{2}$$

1) AR(p)

AR(p) is called an autoregressive process, and its expression is:

$$x_t=\phi_0+\phi_1 x_{t-1}+\phi_2 x_{t-2}+\cdots+\phi_p x_{t-p}+\varepsilon_t \tag{3}$$

Note $\Phi(B)=1-\phi_1 B-\phi_2 B^2-\cdots-\phi_p B^P$,the centralized AR(p) model can be abbreviated as

$$\Phi(B)x_t=\varepsilon_t \tag{4}$$

2) MA(q)

MA(q) is called moving average process, its expression is:

$$x_t=\varepsilon_t+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\cdots+\theta_q\varepsilon_{t-q} \tag{5}$$

if $\Theta(B)=1+\theta_1 B+\theta_2 B^2+\cdots+\theta_q B^P$,the centralized MA(q) model can be abbreviated as

$$x_t=\Theta(B)\,\varepsilon_t \tag{6}$$

3) ARMA (p, q)

ARMA(p, q) is the dependency structure model combining AR(p) and MA(q), and its expression is as follows:

$$x_t=\phi_1 x_{t-1}+\phi_2 x_{t-2}+\cdots+\phi_p x_{t-p}+\theta_1\varepsilon_{t-1}+\theta_2\varepsilon_{t-2}+\cdots+\theta_q\varepsilon_{t-q}+\varepsilon_t \tag{7}$$

$$\Phi(B)x_t=\Theta(B)\,\varepsilon_t \tag{8}$$

4) ARIMA (p, d, q)

After removing the trend by differencing, if

$$\nabla^d x_t \sim ARMA(p, q) \tag{9}$$

Then x_t can be called the differential integrated moving average autoregressive process, and its expressions is

$$x_t \sim ARIMA\ (p, d, q) \tag{10}$$

$$\Phi(B)\nabla^d x_t=\Theta(B)\,\varepsilon_t \tag{11}$$

2.2 Processing ideas

1) Model Identification: That is, to determine the order of the model, that is, to determine the value of p, d, q;

2) Parameter Estimation: Estimate the values of ϕ_1, θ_1, σ_ε^2, μ in the model;

3) Residual Test: Determine whether the model adequately fits the observed data;

4) Model Prediction: Predict future values using a forward forecasting model;

5) Model Comparison: Compare the forecast with the data for 2021.

The specific process of ARIMA modeling is shown in Figure 1

ISSN 2616-5775 Vol. 5, Issue 10: 28-34, DOI: 10.25236/AJCIS.2022.051005

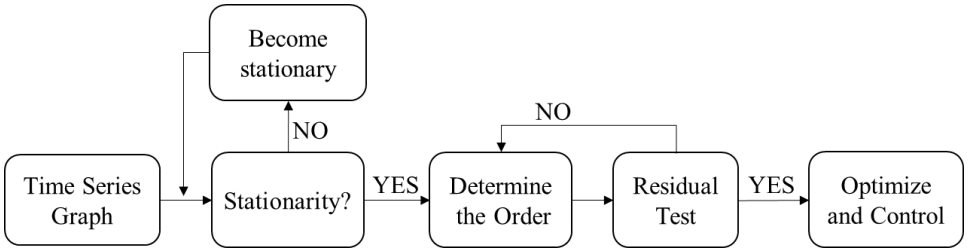

Figure 1: The specific process of ARIMA modeling

3. Modeling Process

3.1 Basic characteristics of data

Since the reform and opening up, Shenzhen's modernization development conforms to the laws of economic development defined by various modern economic theories, conforms to the trend of world modernization development, and at the same time contains Chinese characteristics and wisdom. Compared with other late-developing modernized countries, Shenzhen takes innovation as the core driving force, maintains rapid economic development, and crosses the "middle-income trap" [10]. This paper selects the GDP data of Shenzhen from 1980 to 2020 for modeling, and compress the real data in 2021 with the data predicted by the model to illustrate the accuracy of the model.

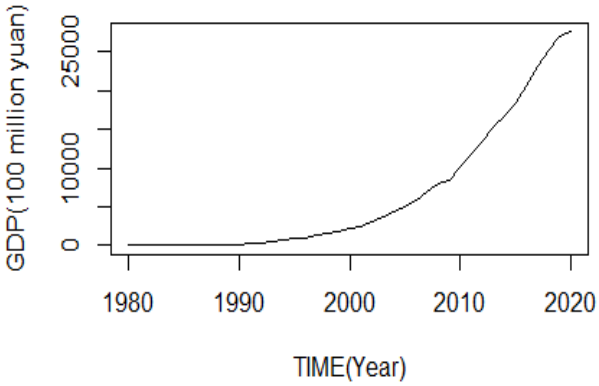

Figure 2: Shenzhen GDP Timing Chart

3.2 Unit root test(ADF test)

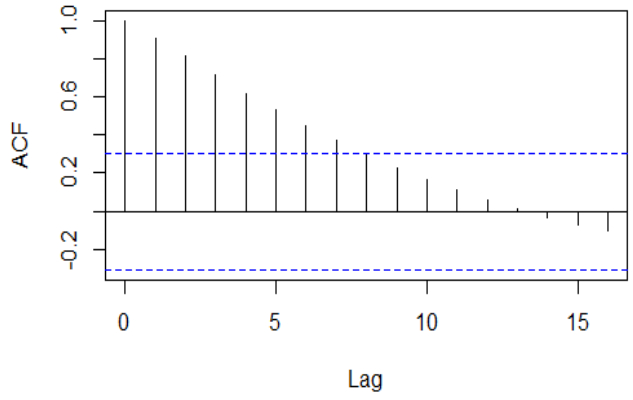

Figure 3: Autocorrelation plot of Shenzhen GDP data

As shown in Figure 2, Shenzhen's GDP has an obvious upward trend, and it can be simply preliminarily judged that the time series is not stable. From the autocorrelation diagram in Figure 3, it can be found that the ACF of the time series decays very slowly, and it can be considered that the time series is non-stationary. In order to more accurately describe the non-stationarity of the series, this paper

ISSN 2616-5775 Vol. 5, Issue 10: 28-34, DOI: 10.25236/AJCIS.2022.051005

considers the use of the ADF test. Therefore, in order to better analyze this data, the data needs to be stabilized. Use R to get that the statistic τ=-1.84 is greater than the corresponding critical value (99% confidence corresponds to -3.98, 95% confidence corresponds to -3.42, 90% confidence corresponds to -3.13). So, accept the null hypothesis that the series is not trend stationary.

In this paper, two methods of Box-Cox transformation and difference operation are used. First perform Box-Cox transformation on the original data, take λ=0, that is, logarithmic change. Second-order difference is then performed on the transformed data. Statistics τ=-4.81, The corresponding critical values are -4.15(99%)，-3.50(95%)，-3.18(90%)，Therefore, the null hypothesis is rejected, and the processed data is trend-stable.

3.3 Build the ARIMA Model

1) Model Identification

The time series diagram, autocorrelation diagram and partial autocorrelation diagram of the preprocessed data are shown in Figure 4, Figure 5 and Figure 6.

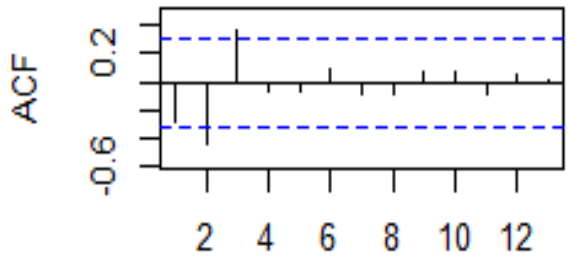

Figure 4: ACF

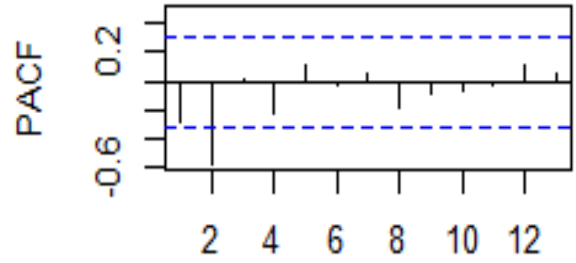

Figure 5: PACF

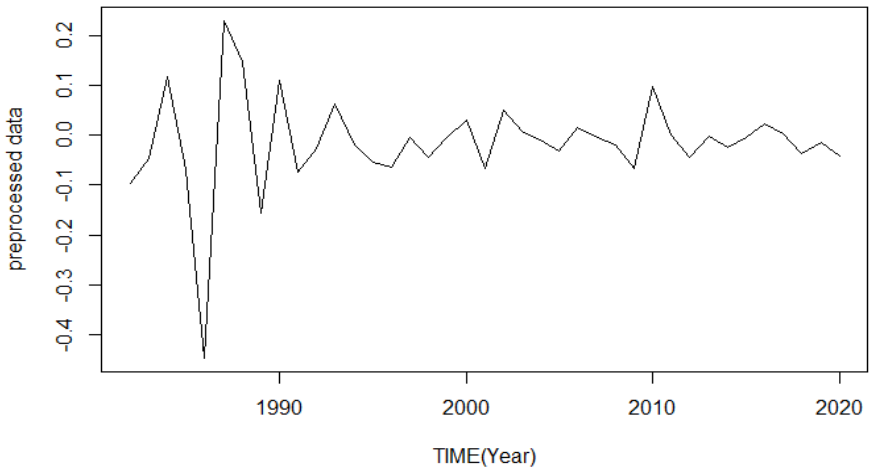

Figure 6: Time series diagram after data processing

Its autocorrelation coefficient is 3-order truncation, and its partial autocorrelation coefficient is 2-order end. For the processed data, consider building the ARIMA(2,0,3) and ARIMA(2,0,2) models, that is, taking the logarithm of the original GDP data to build ARIMA(2, 2, 3) and ARIMA(2, 2, 2)

ISSN 2616-5775 Vol. 5, Issue 10: 28-34, DOI: 10.25236/AJCIS.2022.051005

models. The corresponding AIC, BIC, AIC_c results of the two models are shown in Table 1.

Table 1: Internal comparison results of models

	AIC	BIC	AIC_c
ARIMA(2, 2, 2)	-75.79	-67.47	-73.97
ARIMA(2, 2, 3)	-76.60	-67.62	-73.97

According to the AIC and BIC criteria, the smaller the value is, the smaller the value of the penalty item is, so the model fits better, so the ARIMA(2, 2, 3) model is selected.

2) Parameter Estimation

This paper adopts the method of maximum likelihood estimation to estimate the parameters in the model. Maximum likelihood estimation determines parameter sizes by maximizing the likelihood function. The white noise sequence ε_t in the likelihood function can be represented recursively as a nonlinear function of historical observations and parameters, so that the likelihood function can be determined. To find the maximum value of the likelihood function, a suitable initial value is needed for numerical iteration, so this paper selects the result of least squares estimation as the initial value of maximum likelihood estimation for iteration. The final results are shown in Table 2.

Table 2: Parameter value

Parameter	ϕ_1	ϕ_2	θ_1	θ_2	θ_3
value	0.56	0.40	-1.07	-0.73	0.86
s.e.	0.22	0.19	0.20	0.25	0.18

The parameter's test statistic is $t=\frac{\hat{\beta}}{s.e.}$. If it is greater than 2, the parameter is significant, otherwise it is not significant. It can be seen from Table 1 that each parameter is significant, so the expression of the model is:

$$\left(1-0.56B-0.40B^2\right)(y_t-2y_{t-1}-y_{t-2})+(1+1.07B+0.73B^2-0.86B^3)\varepsilon_t=0$$

After simplification:

$$y_t=2.56y_{t-1}+0.28y_{t-2}-1.36y_{t-3}-0.40y_{t-4}-\varepsilon_t-1.07\varepsilon_{t-1}-0.73\varepsilon_{t-2}+0.86\varepsilon_{t-3}$$

Where $y_t=\ln x_t$, $\varepsilon_t \sim N(0,0.0075)$

3) Residual Test

After setting and estimating a model, it is necessary to test whether the model adequately fits the data and tests the residuals of the model.

$$\hat{\varepsilon_t}=x_t-\hat{x_t} \tag{12}$$

Where $\hat{x_t}=\hat{x}_{t-1}(1)$ is the one-step forward prediction based on the estimated model at time t-1, which is calculated in (4) given.

In this paper, a mixture test is carried out on the residual series. The specific Box-Ljung test statistic is:

$$LB(m)=T(T+2)\sum_{k=1}^{m}\frac{\hat{\rho}_k^2}{T-k} \tag{13}$$

Where $\hat{\rho}_k$ is the ACF of delayed k order samples.

Use the Box. Test function in the R studio to get the test result p=0.9774, the result is significant, indicating that the model makes full use of the given data.

4) h Step Forward Prediction

Set $\mathcal{F}_n$ as historical information. It includes all the information up to n time. Observations x_1, x_2, ···, x_n and random perturbation terms ε_1, ε_2,···,ε_n are regarded as known information. Predict x_{n+h} based on known information. The starting point of the prediction is n and h is the step size of the prediction. So $x_{n+h}=\hat{x_n}(h)$ is a function of historical information, let's call it $f(\mathcal{F}_n)$. In order to measure the accuracy of the forecast, the following mean square forecast error is used to judge.

$$K=E\left(x_{n+h}-f(\mathcal{F}_n)\right)^2 \tag{14}$$

ISSN 2616-5775 Vol. 5, Issue 10: 28-34, DOI: 10.25236/AJCIS.2022.051005

Calculated that $K \geq E(x_{n+h} - E(x_{n+h}|\mathcal{F}_n))^2$

The above equal sign holds if and only if $f(F_n) = E(x_{n+h}|\mathcal{F}_n)$, that is

$$\operatorname{argmin}_f K = E(x_{n+h}|\mathcal{F}_n) \tag{15}$$

That is, the optimal mean square error prediction is the conditional expectation of the prediction target under the given historical information. Based on the above principles, the GDP of Shenzhen from 2021 to 2025 is predicted, and the results are shown in Table 3 and Figure 7.

Considering the national bond interest rate of 1.5% in 2020, selecting the GDP of the previous year as the base period, and calculating the comparable price in the next year, the corresponding year-on-year growth rate can be obtained, as shown in Table 3

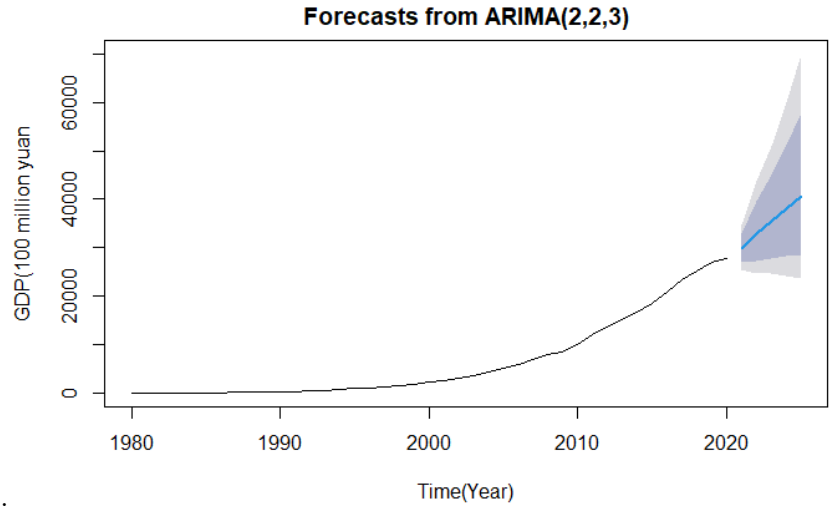

Figure 7: Prediction results

5) Analysis of results

In 2021, Shenzhen's actual GDP is 3,066.485 billion yuan, and the error is a relative error of 2.9%. The fluctuation of the relative error is small. The predicted year-on-year growth rate in 2021 is 6.1%, which is close to the actual 6.7%. The prediction results are relatively stable and have certain reference value. It is predicted that the GDP of Shenzhen from 2022 to 2025 will be 3280.178 billion yuan, 3538.896 billion yuan, 3808.253 billion yuan, and 4050.760 billion yuan respectively. The results show that Shenzhen will break through the 4 trillion-yuan GDP mark in 2021-2025, that is, during the "14th Five-Year Plan" period. The projected GDP growth rate will remain between 5% and 8%. However, due to the impact of COVID-19, although Shenzhen's GDP in the first quarter of 2022 exceeded 700 billion yuan, ranking first in Guangdong Province, with a year-on-year increase of 2%, there is still a certain distance from the predicted goal. From the interval forecast results from 2023 to 2025, it can be seen that the interval span is large, indicating that the model has large uncertainty and general accuracy for long-term forecasting. Therefore, the ARIMA (2, 2, 3) model is more suitable for short-term forecasting. For long-term forecasts, due to the existence of unknown variables such as economy, culture, geopolitics, etc., it is difficult for the model to estimate the data more accurately.

Table 3: Forecast results, confidence interval and forecast growth rate (Unit: 100 million yuan)

year	point forecast	80% confidence interval	95% confidence interval	growth rate
2021	29799.40	(26929.58，32975.04)	(25524.02，34790.91)	6.10%
2022	32801.78	(27236.09，39504.82)	(24682.94，43591.11)	8.45%
2023	35388.96	(27870.40，44935.78)	(24560.36，50991.85)	6.30%
2024	38082.53	(28294.30，51256.93)	(24176.66，59986.74)	6.02%
2025	40597.69	(28412.93，58007.84)	(23521.86，70096.82)	5.03%

4. Conclusion

This paper analyzes the GDP data of Shenzhen from 1980 to 2020, and predicts the GDP data of

ISSN 2616-5775 Vol. 5, Issue 10: 28-34, DOI: 10.25236/AJCIS.2022.051005

Shenzhen in the next five years by establishing the ARIMA (2, 2, 3) model. First, logarithmic transformation and second-order difference transformation were performed on the data to convert the original sequence into a stationary sequence. Select the order from the results of the ACF and PACF plots. According to the "AIC" and "BIC" criteria of model order, the residuals passed the mixed test, and finally determined the ARIMA (2, 2, 3) model. The ARIMA model is used to obtain the forecast data of Shenzhen's GDP for 5 years from 2021 to 2025, and the confidence intervals of 80% and 95% are given, and a relative error of 2.9% between the predicted results and the actual results in 2021 is obtained. It must be pointed out that the "ARIMA" model only predicts from historical data, and does not take into account the political, economic, and emergencies that may occur in the future. From a long-term perspective, it is difficult for the model to predict completely and accurately, but the model still has a certain reference value in short-term prediction.

References

[1] Guo Zhiwu, Pu Jihong, Teng Guozhao. Study on the Method of Chinese New Year Factor's Adjustment Based on ARIMA model [J]. China Health Statistics, 2009,26 (6):573-576,579.

[2] Zheng Huimin, Xue Yunlian, Huang Yanfei, et al. Application of ARIMA model to predicting the incidence tendency of notifiable communicable diseases in Shenzhen City [J]. Practical Preventive Medicine, 2016, 23(2): 240-243.

[3] Yan Zhouning, Mu Jingfeng, Zhao Xing, et al. The time series prediction of PM2.5 in Shenzhen based on ARIMA model [J]. Modern Preventive Medicine, 2018, 45(2):220-223,242.

[4] Yi Zhiguo, Xiang Lili. Research on cigarette sales forecast based on ARIMA model [J]. Chinese and Foreign Entrepreneurs, 2021(11):149.

[5] Chen Congcong. Prediction and Analysis of Shandong Province's GDP Based on ARIMA Model and ARIMAX Model [D]. Shandong University, 2016.

[6] Yan Yanwen. Analysis and forecast of GDP in Shandong Province Based on ARIMA Model [J]. Mathematics in Practice and Theory, 2018,48(04):285-292.

[7] Zhao Zimeng. Predicting Chengdu's GDP based on ARIMA time series model [J]. Communication World, 2019, 26(02): 206-207.

[8] Xu Mingyan. Prediction and Analysis of Jiangsu Province's GDP Based on ARIMA Model and BP Neural Network Model [D]. Shandong University, 2020.

[9] Qu Haiqing, He Xianping. Research on GDP forecasting model of Hubei Province based on time series analysis [J]. Journal of Hubei University of Economics (Humanities and Social Sciences), 2021, 18(09): 37-39.

[10] Xie Zhikui, Li Zhuo. Shenzhen Model: World Trend and Chinese Characteristics: A Theoretical Interpretation of the Achievements of Shenzhen's Modernization from 1978 [J]. Shenzhen Social Sciences, 2019(01): 97-110+159